

## Week 8 – Problem Set

1.

Economists are often concerned about using survey data—i.e., what individuals *say* they would do rather than what those individuals *actually* do—because survey respondents can be overly optimistic or pessimistic when the truth is complex (“How much more would your business invest under the tax plan of Candidate A?”), or deliberately untruthful when the truth is embarrassing (“When was the last time you cheated on your spouse?”). Consider the simple population model  $y_i = \beta_0 + \beta_1 x_{1i}^* + u_{it}$  where  $x_{1i}^*$  is unobserved and  $x_{1i} = x_{1i}^* + r_i$  is a measure of  $x_{1i}^*$  generated from survey data. Under what circumstances would you find the Classical Errors-in-Variables (CEV) assumption  $cov[x_{1i}^*, r_i] = 0$  to be reasonable?

2.

Kiel and McClain (1995) used the pooled cross section data set `hprice.txt`—which contains the prices and characteristics of homes sold in North Andover, Massachusetts in 1978 and 1981—to measure the relationship between price and the distance to a newly-built garbage incinerator. An extensive public conversation concerning the incinerator began in 1979 and its construction began in 1981. The variables are listed on the following page.

(a) Are these data independent and identically distributed (i.i.d.)? Why or why not?

(b) Using the data for year=1981 only, estimate the linear equation

$$\log(\text{price}_{i,1981}) = \beta_0 + \beta_1 \log(\text{dist}_{i,1981}) + u_{i,1981}$$

and explain whether it is appropriate for determining the causal effects of proximity to the incinerator on housing prices.

(c) Using both years of data, estimate the difference-in-differences regression

$$\log(\text{price}_{it}) = \delta_0 + \delta_1 1[\text{year} = 1981] + \delta_2 \log(\text{dist}_{it}) + \delta_3 1[\text{year} = 1981] \cdot \log(\text{dist}_{it}) + u_{it}$$

where  $1[\text{year} = 1981]$  is a dummy equal to 1 if  $\text{year} = 1981$ . Perform a robust test of the null hypothesis that the building incinerator had no effect on home prices. Discuss results.

(d) Is the equation in (c) a structural model? Why or why not? Why would it matter?

(e) Repeat the test in (b) adding the variables  $\log(\text{intst})$ ,  $(\log(\text{intst}))^2$ ,  $\log(\text{area})$ ,  $\log(\text{land})$ ,  $\text{age}$ ,  $\text{age}^2$ ,  $\text{rooms}$ , and  $\text{baths}$  to the regression. Does the addition of these regressors change your interpretation of the economic meaning of  $\hat{\delta}_3$ ?

<u>Variable Name</u>	<u>Variable Label</u>
year	1978 or 1981
age	age of house
nbh	neighborhood indicator (1-6)
cbd	distance from house to central business district (in feet)
intst	distance from house to interstate (in feet)
price	selling price
rooms	number of rooms in house
area	square footage of house
land	square footage lot
baths	number of bathrooms
dist	distance from house to incinerator (in feet)

3.

This problem asks you to consider difference-in-differences estimation when panel data are available. Consider the following regression as a means of estimating the effect of a policy change with  $T = 2$ :

$$y_{it} = \theta_1 + \theta_2 1[t = 2] + \delta_1 \text{prog}_{it} + c_i + u_{it}$$

where  $y$  is the outcome of interest,  $1[t = 2]$  is a dummy equal to 1 if  $t = 2$ ,  $\text{prog}_{i1} = 0$ ,  $\text{prog}_{i2}$  is a dummy equal to 1 if  $i$  is affected by the policy change in  $t = 2$ , and  $E[u_{it} | \text{prog}_{i2}, c_i] = 0$ .

(a) Why is it important to include  $c_i$  and  $1[t = 2]$  in the equation?

(b) Show that  $\hat{\delta}_1 = \overline{\Delta y}_{\text{treatment}} - \overline{\Delta y}_{\text{control}}$  and is therefore the difference-in-differences estimator of the effect of the policy change.

4.

The panel data set `ezunem.txt` provides data on unemployment claims in 22 cities from 1980-1988 and when each city introduced an enterprise zone for economic development. The variables include:

<u>Variable Name</u>	<u>Variable Label</u>
<code>city</code>	city identifier
<code>year</code>	1980-1988
<code>uclms</code>	number of unemployment claims
<code>ez</code>	=1 if the city has an enterprise zone

where `city` is the cross section  $i$  identifier and `year` is the time  $t$  identifier.

(a) Consider the random growth model

$$\log(\text{uclms}_{it}) = \theta_t + c_i + g_i \cdot t + \beta_1 \text{ez}_{it} + u_{it}$$

What are the economic interpretations of  $g_i$  and  $\beta_1$ ?

(b) Write down the estimating equation that results if one first-differences the structural equation in (a) and then applies the fixed effects transformation on the differenced equation.

(c) Estimate the model in (a) with robust standard errors using the two-step method described in (b). Discuss results.

(d) Consider the random growth model

$$\log(\text{uclms}_{it}) = \theta_t + c_i + g_i \cdot t + \beta_1 \text{ez}_{it} + \beta_2 \text{ez}_{it} \cdot t + u_{it}$$

Write down the estimating equation that results if one first-differences the structural equation and then applies the fixed effects transformation on the differenced equation. (Hint: Be careful with the treatment of  $\text{ez}_{it} \cdot t$ ; the time trend does not simply disappear.)

(e) Estimate the model in (d) with robust standard errors using the two-step method. Discuss results.

5.

As a general matter, the Bootstrap Method is most useful when obtaining the standard error of interest would otherwise require multiple asymptotic approximations—say, when finding the standard error of an average partial effect from the estimates of a nonlinear model using the Delta Method. It is unnecessary in situations that do not require an asymptotic approximation, such as when the measure of interest is a parameter (or linear function of parameters) from a linear model. Nonetheless, it is useful to first learn the Bootstrap Method using a simple linear model because the code can be easily adjusted to handle the more interesting scenarios.

In Problem Set 2, Problem 4 you reported the OLS estimates and robust (asymptotic) standard errors for the simple model

$$stndfnl_i = \beta_0 + \beta_1 atndrte_i + \beta_2 frosh_i + \beta_3 soph_i + u_i$$

This week I have provided a SAS “macro” that performs both a *paired* (or *nonparametric*) and a *wild* bootstrap for this model. Specifically, the macro allows you to specify the type (paired or wild), the sample size  $n^*$ , and the number of bootstrap iterations  $M$ . The macro outputs a table of the parameter estimates, the usual asymptotic results (i.e., standard errors, t-statistics, and p-values), and the bootstrap results (i.e., biases, standard errors, t-statistics, and critical t-values). Using the lecture slides on the Bootstrap Method as a reference, read through the comments in the code and make sure that you can follow each step.

(a) Execute the macro to perform a paired bootstrap using a bootstrap sample size of  $n^* = 680$  (which is the sample size of the original data) and  $M = 100$  iterations. Does the significance of any of your parameter estimates change using the bootstrap standard errors as opposed to the asymptotic standard errors?

(b) Execute the macro again to perform a paired bootstrap using the same values  $n^* = 680$  and  $M = 100$ . Do you obtain the same standard errors and critical values as you did in (a)? Why or why not?

(c) Using  $n^* = 680$ , perform a paired bootstrap for  $M = 500$  and  $M = 1000$ . Do the bootstrap biases and standard errors change in a systematic way as  $M$  increases?

(d) As we discussed in Lecture 8, the wild bootstrap incorporates heteroskedasticity in the error term and is used for robust standard errors. Specifically, the error term is expressed as  $f(\hat{u}) \cdot v$  for some function  $f(\cdot)$  and some zero-mean distribution  $f(v)$ : the first term incorporates the heteroskedasticity (because  $\hat{u}_i = y_i - \mathbf{x}_i\hat{\beta}$  is obviously a function of  $\mathbf{x}$ ) and the second term ensures that error term has a conditional expectation of zero. The bootstrap macro uses the most common choice for the function  $f(\hat{u})$  and for the distribution  $f(v)$ . Examining the macro, what is  $f(\hat{u})$  and  $f(v)$ ?

(e) Using  $n^* = 680$ , perform a wild bootstrap for  $M = 100$ ,  $M = 500$ , and  $M = 1000$ . How do your results compare with your findings in (a) and (c)?